METHODS FOR RAPID DETECTION AND IDENTIFICATION OF BIOAGENTS IN EPIDEMIOLOGICAL AND FORENSIC INVESTIGATIONS

CROSS-REFERENCE TO RELATED APPLICATIONS

This application is a continuation-in-part of U.S. application Serial No. 10/323,438 filed December 18, 2002, which is incorporated herein by reference in its entirety. This application is also a continuation-in-part of U.S. application Serial No. 09/798,007 filed March 2, 2001, which is incorporated herein by reference in its entirety. This application also claims priority to U.S. provisional application Serial No. 60/431,319 filed December 6, 2002 and to U.S. provisional application Serial No. 60/461,494 filed April 9, 2003, each of which is incorporated herein by reference in its entirety.

STATEMENT OF GOVERNMENT SUPPORT

This invention was made with United States Government support under DARPA/SPO contract BAA00-09. The United States Government may have certain rights in the invention.

FIELD OF THE INVENTION

This invention relates to the field of forensic and epidemiological investigations. The methods provide rapid identification of known or suspected terrorists or criminals based on forensic evidence. Additionally, the methods provide tracking of the geographic locations of the terrorists or criminals by microbial geographic profiling of bioagents associated with the terrorists or criminals. Furthermore, the methods provide genotyping of bioagents such as those associated with acts of biowarfare, terrorism or criminal activity.

BACKGROUND OF THE INVENTION

The ease with which small countries and terrorist groups can now obtain biological warfare agents has escalated the need to provide the war fighter and civilians alike with miniature, easy to use, disposable instruments for detection and identification of potentially hazardous biological agents (Iqbal *et al. Biosensors & Bioelectronics* **2000**, *15*, 549-578; Christel, L.A., *et al. J. Biomech. Eng.* **1999**, *121*, 22–27; Higgins, J.A., *et al. Ann. NY Acad. Sci.* 1999, *894*, 130–148; and Hood, E. *Environ. Health Perspect.* **1999**, *107*, 931–932). Traditional methods for detection and identification of microorganisms, viruses and/or their products lack the speed and sensitivity to be of field usage since they are not real time or even typically completed in a single day. Microbial and viral identification assays have as their basis, the principle that dates back to the days of Pasteur, i.e. the growth of the organism in culture or replication of virus in a suitable host (Reischl, U., *Frontiers Biosci.,* **1996**, *1,* Application of molecular biology-based methods to the diagnosis of infectious diseases.. 1, e72–e77). Toxin identification has typically relied on biological assays, which although relatively rapid, usually requires purification of the toxin prior to testing (Feng, P. *Mol. Biotechnol.* **1997**, *7*, 267–278; van der Zee, H. *et al. J. AOAC Int.* **1997,** *80*, 934–940). Depending upon the nature of the agent to be detected, this process can take from days to months (Pillai, S.D. *Arch. Virol.* **1997**, *13* Suppl., 67–82; van der Zee, H. *et al. J. AOAC Int.* **1997,** *80*, 934–940). Clearly, this is not practical in situations where detection and identification may be required for protecting a population from hazardous biological agents. Molecular recognition systems that can be used for rapid identification can improve response time and thus avert or reduce the number of casualties associated with a potential bioterrorism or biowarfare event.

Biological threat agents can be either infectious or toxigenic organisms or simply toxins (Hood, E. *Environ. Health Perspect.* **1999**, *107*, 931–932; Haines, J.D. *et al J. Okla. State Med. Assoc.* **1999**, *93,* 187–196). Examples of the former are *Bacillus anthracis* (anthrax) and *Yersinia pestis* (plague) while the latter is exemplified by staphylococcal enterotoxin B or botulinum toxin. Detection of toxins has followed a similar track as that observed for detection of chemical threat agents. Typically, toxins are detected on the basis of their respective chemical structures. Detection of mycotoxins has been accomplished using traditional analytical chemistry tools such as gas chromatography-mass spectrometry (GC-MS) (Black, R.M *et al. J. Chromatogr.* **1986**, *367*, 103–116). Although precise and highly sensitive, GC-MS does not lend itself to field applications and cannot be easily applied to complex target analytes such as bacteria. Specific compounds, i.e. signature components might be identified in targeted bacterial agents but this approach tends to be too complex for routine, high throughput analysis (van der Zee, H. *et al. J. AOAC Int.* **1997,** *80*, 934–940; Hood, E. *Environ. Health Perspect.* **1999,** *107*, 931–932).

A number of technological innovations have provided tools that have made detection and identification of microorganisms, viruses and their products faster and more sensitive. Two significant technologies that have had dramatic impact on potentially rapid detection are: (1) generation of monoclonal antibodies; and (2) collection of individual methodological advances that have formed the basis of recombinant DNA technology. A key event in the latter technology is the development of polymerase chain reaction (PCR) technology which as a singular process, has significantly reshaped the authors' thinking with respect to detection of biological agents.

Identification of biological threat agents involves recognition of bacteria (vegetative cells and spores), viruses and toxins. Nucleic acid and immunology-based methods for identification of bacteria, viruses and their products (antigens and toxins) have found wide application including testing of food, clinical and environmental samples. Extension of these applications for detection and identification of biological threat agents is reasonable since basic principles involved are identical. Two essential features characterize these applications, i.e. the need to: (1) develop a target specific identification method; and (2) formulate an assay that will work on the requisite sample. However, the majority of the methods developed meet the first criterion, i.e. identify the target organism or toxin, but fail to be sufficiently robust to work on “real” samples.

Mass spectrometry provides detailed information about the molecules being analyzed, including high mass accuracy. It is also a process that can be easily automated. Low-resolution MS may be unreliable when used to detect some known agents, if their spectral lines are sufficiently weak or sufficiently close to those from other living organisms in the sample. DNA chips with specific probes can only determine the presence or absence of specifically anticipated organisms. Because there are hundreds of thousands of species of benign bacteria, some very similar in sequence to threat organisms, even arrays with 10,000 probes lack the breadth needed to detect a particular organism.

Antibodies face more severe diversity limitations than arrays. If antibodies are designed against highly conserved targets to increase diversity, the false alarm problem will dominate, again because threat organisms are very similar to benign ones. Antibodies are only capable of detecting known agents in relatively uncluttered environments.

Several groups have described detection of PCR products using high resolution electrospray ionization-Fourier transform-ion cyclotron resonance mass spectrometry (ESI-FT-ICR MS). Accurate measurement of exact mass combined with knowledge of the number of at least one nucleotide allowed calculation of the total base composition for PCR duplex products of approximately 100 base pairs. (Aaserud *et al.*, *J. Am. Soc. Mass Spec.*, 1996, *7*, 1266-1269; Muddiman *et al.*, *Anal. Chem.*, 1997, *69*, 1543-1549; Wunschel *et al.*, A*nal. Chem.*, 1998, *70*, 1203-1207; Muddiman *et al.*, *Rev. Anal. Chem.*, 1998, *17*, 1-68). Electrospray ionization-Fourier transform-ion cyclotron resistance (ESI-FT-ICR) MS may be used to determine the mass of double-stranded, 500 base-pair PCR products via the average molecular mass (Hurst *et al.*, *Rapid Commun. Mass Spec.* 1996, *10*, 377-382). The use of matrix-assisted laser desorption ionization-time of flight (MALDI-TOF) mass spectrometry for characterization of PCR products has been described. (Muddiman *et al.*, *Rapid Commun. Mass Spec.*, 1999, *13*, 1201-1204). However, the degradation of DNAs over about 75 nucleotides observed with MALDI limited the utility of this method.

U.S. Patent No. 5,849,492 describes a method for retrieval of phylogenetically informative DNA sequences which comprise searching for a highly divergent segment of genomic DNA surrounded by two highly conserved segments, designing the universal primers for PCR amplification of the highly divergent region, amplifying the genomic DNA by PCR technique using universal primers, and then sequencing the gene to determine the identity of the organism.

U.S. Patent No. 5,965,363 discloses methods for screening nucleic acids for polymorphisms by analyzing amplified target nucleic acids using mass spectrometric techniques and to procedures for improving mass resolution and mass accuracy of these methods.

WO 99/14375 describes methods, PCR primers and kits for use in analyzing preselected DNA tandem nucleotide repeat alleles by mass spectrometry.

WO 98/12355 discloses methods of determining the mass of a target nucleic acid by mass spectrometric analysis, by cleaving the target nucleic acid to reduce its length, making the target single-stranded and using MS to determine the mass of the single-stranded shortened target. Also disclosed are methods of preparing a double-stranded target nucleic acid for MS analysis comprising amplification of the target nucleic acid, binding one of the strands to a solid support, releasing the second strand and then releasing the first strand which is then analyzed by MS. Kits for target nucleic acid preparation are also provided.

PCT WO97/33000 discloses methods for detecting mutations in a target nucleic acid by nonrandomly fragmenting the target into a set of single-stranded nonrandom length fragments and determining their masses by MS.

U.S. Patent No. 5,605,798 describes a fast and highly accurate mass spectrometer-based process for detecting the presence of a particular nucleic acid in a biological sample for diagnostic purposes.

WO 98/21066 describes processes for determining the sequence of a particular target nucleic acid by mass spectrometry. Processes for detecting a target nucleic acid present in a biological sample by PCR amplification and mass spectrometry detection are disclosed, as are methods for detecting a target nucleic acid in a sample by amplifying the target with primers that contain restriction sites and tags, extending and cleaving the amplified nucleic acid, and detecting the presence of extended product, wherein the presence of a DNA fragment of a mass different from wild-type is indicative of a mutation. Methods of sequencing a nucleic acid via mass spectrometry methods are also described.

WO 97/37041, WO 99/31278 and U.S. Patent No. 5,547,835 describe methods of sequencing nucleic acids using mass spectrometry. U.S. Patent Nos. 5,622,824, 5,872,003 and 5,691,141 describe methods, systems and kits for exonuclease-mediated mass spectrometric sequencing.

Thus, there is a need for methods for bioagent detection and identification which is both specific and rapid, and in which no nucleic acid sequencing is required. The present invention addresses this need as well as other needs.

SUMMARY OF THE INVENTION

The present invention is directed to methods of identification of a bioagent associated with an act of biowarfare, terrorism or criminal activity by determining a first molecular mass of a first amplification product of a first bioagent identifying amplicon obtained from a sample taken at the scene of biowarfare, terrorism or criminal activity and comparing the first molecular mass to a second molecular mass of a second bioagent identifying amplicon wherein both first and second bioagent identifying amplicons are correlative.

The present invention is also directed to forensic methods for tracking the geographic location of a bioagent associated with an act of biowarfare by determining a first molecular mass of a first amplification product of a first bioagent identifying amplicon from a forensic sample obtained from a given geographic location; and comparing the first molecular mass to a second molecular mass of a second bioagent identifying amplicon wherein both first and second bioagent identifying amplicons are correlative, wherein a match between the first molecular mass and the second molecular mass indicates at least transient presence of the bioagent associated with an act of biowarfare at the given geographic location.

The present invention is also directed to methods of genotyping a bioagent, by determining a first molecular mass of a first amplification product of a first bioagent identifying amplicon that contains genotyping information and comparing the first molecular mass to a second molecular mass of a second bioagent identifying amplicon that contains genotyping information, wherein the first and second bioagent identifying amplicons are correlative, and wherein a match between the first molecular mass and the second molecular mass identifies a genotype of the bioagent.

The present invention is also directed to methods of tracking a known or suspected terrorist or criminal by determining a first molecular mass of a first amplification product of a first bioagent identifying amplicon containing microbial geographic profiling information from a forensic sample known to be associated with the terrorist or criminal and comparing the first molecular mass to a second molecular mass of a second bioagent identifying amplicon wherein the second bioagent identifying amplicon contains microbial geographic profiling information, wherein both first and second bioagent identifying amplicons are correlative, and wherein a match between the first molecular mass and the second molecular mass indicates at least transient presence of the known or suspected criminal at the geographic location indicated by the microbial geographic profiling information.

BRIEF DESCRIPTION OF THE DRAWINGS

Figures 1A-1H and Figure 2 are consensus diagrams that show examples of conserved regions from 16S rRNA (Fig. 1A-1, 1A-2, 1A-3, 1A-4, and 1A-5), 23S rRNA (3'-half, Fig. 1B, 1C, and 1D; 5'-half, Fig. 1E-F), 23S rRNA Domain I (Fig. 1G), 23S rRNA Domain IV (Fig. 1H) and 16S rRNA Domain III (Fig. 2) which are suitable for use in the present invention. Lines with arrows are examples of regions to which intelligent primer pairs for PCR are designed. The label for each primer pair represents the starting and ending base number of the amplified region on the consensus diagram. Bases in capital letters are greater than 95% conserved; bases in lower case letters are 90-95% conserved, filled circles are 80-

90% conserved; and open circles are less than 80% conserved. The label for each primer pair represents the starting and ending base number of the amplified region on the consensus diagram. The nucleotide sequence of the 16S rRNA consensus sequence is SEQ ID NO:3 and the nucleotide sequence of the 23S rRNA consensus sequence is SEQ ID NO:4.

Figure 2 shows a typical primer amplified region from the 16S rRNA Domain III shown in Figure 1A-1.

Figure 3 is a schematic diagram showing conserved regions in RNase P. Bases in capital letters are greater than 90% conserved; bases in lower case letters are 80-90% conserved; filled circles designate bases which are 70-80% conserved; and open circles designate bases that are less than 70% conserved.

Figure 4 is a schematic diagram of base composition signature determination using nucleotide analog "tags" to determine base composition signatures.

Figure 5 shows the deconvoluted mass spectra of a *Bacillus anthracis* region with and without the mass tag phosphorothioate A (A*). The two spectra differ in that the measured molecular weight of the mass tag-containing sequence is greater than the unmodified sequence.

Figure 6 shows base composition signature (BCS) spectra from PCR products from *Staphylococcus aureus* (*S. aureus* 16S_1337F) and *Bacillus anthracis* (*B. anthr.* 16S_1337F), amplified using the same primers. The two strands differ by only two (AT-->CG) substitutions and are clearly distinguished on the basis of their BCS.

Figure 7 shows that a single difference between two sequences (A14 in *B. anthracis* vs. A15 in *B. cereus*) can be easily detected using ESI-TOF mass spectrometry.

Figure 8 is an ESI-TOF of *Bacillus anthracis* spore coat protein sspE 56mer plus calibrant. The signals unambiguously identify *B. anthracis* versus other Bacillus species.

Figure 9 is an ESI-TOF of a *B. anthracis* synthetic 16S_1228 duplex (reverse and forward strands). The technique easily distinguishes between the forward and reverse strands.

Figure 10 is an ESI-FTICR-MS of a synthetic *B. anthracis* 16S_1337 46 base pair duplex.

Figure 11 is an ESI-TOF-MS of a 56mer oligonucleotide (3 scans) from the *B. anthracis* saspB gene with an internal mass standard. The internal mass standards are designated by asterisks.

Figure 12 is an ESI-TOF-MS of an internal standard with 5 mM TBA-TFA buffer showing that charge stripping with tributylammonium trifluoroacetate reduces the most abundant charge state from [M-8H+]8- to [M-3H+]3-.

Figure 13 is a portion of a secondary structure defining database according to one embodiment of the present invention, where two examples of selected sequences are displayed graphically thereunder.

Figure 14 is a three dimensional graph demonstrating the grouping of sample molecular weight according to species.

Figure 15 is a three dimensional graph demonstrating the grouping of sample molecular weights according to species of virus and mammal infected.

Figure 16 is a three dimensional graph demonstrating the grouping of sample molecular weights according to species of virus, and animal-origin of infectious agent.

Figure 17 is a figure depicting how the triangulation method of the present invention provides for the identification of an unknown bioagent without prior knowledge of the unknown agent. The use of different primer sets to distinguish and identify the unknown is also depicted as primer sets I, II and III within this figure. A three dimensional graph depicts all of bioagent space (**170**), including the unknown bioagent, which after use of primer set I (**171**) according to a method according to the present invention further differentiates and classifies bioagents according to major classifications (**176**) which, upon further analysis using primer set II (**172**) differentiates the unknown agent (**177**) from other, known agents (**173**) and finally, the use of a third primer set (**175**) further specifies subgroups within the family of the unknown (**174**).

Figure 18 depicts three representative different mass spectral traces of bioagent identifying amplicons, each containing a single nucleotide polymorphism, indicating that molecular mass or base composition signature is capable of distinguishing the single nucleotide polymorphism.

DESCRIPTION OF EMBODIMENTS

A. Introduction

The present invention provides, *inter alia*, methods for detection and identification of bioagents in an unbiased manner using "bioagent identifying amplicons." "Intelligent primers" are selected to hybridize to conserved sequence regions of nucleic acids derived from a bioagent and which bracket variable sequence regions to yield a bioagent identifying amplicon which can be amplified and which is amenable to molecular mass determination. The molecular mass then provides a means to uniquely identify the bioagent without a requirement for prior knowledge of the possible identity of the bioagent. The molecular mass or corresponding "base composition signature" (BCS) of the amplification product is then matched against a database of molecular masses or base composition signatures. Furthermore, the method can be applied to rapid parallel "multiplex" analyses, the results of which can be employed in a triangulation identification strategy. The present method provides rapid throughput and does not require nucleic acid sequencing of the amplified target sequence for bioagent detection and identification.

B. Bioagents

In the context of this invention, a "bioagent" is any organism, cell, or virus, living or dead, or a nucleic acid derived from such an organism, cell or virus. Examples of bioagents include, but are not limited, to cells, including but not limited to, cells, including but not limited to human clinical samples, bacterial cells and other pathogens) viruses, fungi, and protists, parasites, and pathogenicity markers (including but not limited to: pathogenicity islands, antibiotic resistance genes, virulence factors, toxin genes and other bioregulating compounds). Samples may be alive or dead or in a vegetative state (for example, vegetative bacteria or spores) and may be encapsulated or bioengineered. In the context of this invention, a "pathogen" is a bioagent which causes a disease or disorder.

Despite enormous biological diversity, all forms of life on earth share sets of essential, common features in their genomes. Bacteria, for example have highly conserved sequences in a variety of locations on their genomes. Most notable is the universally conserved region of the ribosome. but there are also conserved elements in other non-coding RNAs, including RNAse P and the signal recognition particle (SRP) among others. Bacteria have a common set of absolutely required genes. About 250 genes are present in all bacterial species (*Proc. Natl. Acad. Sci. U.S.A.*, **1996**, *93*, 10268; *Science*, **1995**, *270*, 397), including tiny genomes like *Mycoplasma*, *Ureaplasma* and *Rickettsia*. These genes encode proteins involved in translation, replication, recombination and repair, transcription, nucleotide metabolism, amino acid metabolism, lipid metabolism, energy generation, uptake, secretion and the like. Examples of these proteins are DNA polymerase III beta, elongation factor TU, heat shock protein groEL, RNA polymerase beta, phosphoglycerate kinase, NADH dehydrogenase, DNA ligase, DNA topoisomerase and elongation factor G. Operons can also be targeted using the present method. One example of an operon is the bfp operon from enteropathogenic *E. coli.* Multiple core chromosomal genes can be used to classify bacteria at a genus or genus species level to determine if an organism has threat potential. The methods can also be used to detect pathogenicity markers (plasmid or chromosomal) and antibiotic resistance genes to confirm the threat potential of an organism and to direct countermeasures.

C. Selection of "Bioagent Identifying Amplicons"

Since genetic data provide the underlying basis for identification of bioagents by the methods of the present invention, it is necessary to select segments of nucleic acids which ideally provide enough variability to distinguish each individual bioagent and whose molecular mass is amenable to molecular mass determination. In one embodiment of the present invention, at least one polynucleotide segment is amplified to facilitate detection and analysis in the process of identifying the bioagent. Thus, the nucleic acid segments which provide enough variability to distinguish each individual bioagent and whose molecular masses are amenable to molecular mass determination are herein described as "bioagent identifying amplicons." The term "amplicon" as used herein, refers to a segment of a polynucleotide which is amplified in an amplification reaction.

As used herein, "intelligent primers" are primers that are designed to bind to highly conserved sequence regions of a bioagent identifying amplicon that flank an intervening variable region and yield amplification products which ideally provide enough variability to distinguish each individual bioagent, and which are amenable to molecular mass analysis. By the term "highly conserved," it is meant that the sequence regions exhibit between about 80-100%, or between about 90-100%, or between about 95-100% identity. The molecular mass of a given amplification product provides a means of identifying the bioagent from which it was obtained, due to the variability of the variable region. Thus design of intelligent primers requires selection of a variable region with appropriate variability to resolve the identity of a given bioagent. Bioagent identifying amplicons are ideally specific to the identity of the bioagent. A plurality of bioagent identifying amplicons selected in parallel for distinct bioagents which contain the same conserved sequences for hybridization of the same pair of intelligent primers are herein defined as "correlative bioagent identifying amplicons."

In one embodiment, the bioagent identifying amplicon is a portion of a ribosomal RNA (rRNA) gene sequence. With the complete sequences of many of the smallest microbial genomes now available, it is possible to identify a set of genes that defines "minimal life" and identify composition signatures that uniquely identify each gene and organism. Genes that encode core life functions such as DNA replication, transcription, ribosome structure, translation, and transport are distributed broadly in the bacterial genome and are suitable regions for selection of bioagent identifying amplicons. Ribosomal RNA (rRNA) genes comprise regions that provide useful base composition signatures. Like many genes involved in core life functions, rRNA genes contain sequences that are extraordinarily conserved across bacterial domains interspersed with regions of high variability that are more specific to each species. The variable regions can be utilized to build a database of base composition signatures. The strategy involves creating a structure-based alignment of sequences of the small (16S) and the large (23S) subunits of the rRNA genes. For example, there are currently over 13,000 sequences in the ribosomal RNA database that has been created and maintained by Robin Gutell, University of Texas at Austin, and is publicly available on the Institute for Cellular and Molecular Biology web page on the world wide web of the Internet at, for example, "rna.icmb.utexas.edu/." There is also a publicly available rRNA database created and maintained by the University of Antwerp, Belgium on the world wide web of the Internet at, for example, "rrna.uia.ac.be."

These databases have been analyzed to determine regions that are useful as bioagent identifying amplicons. The characteristics of such regions include: a) between about 80 and 100%, or greater than about 95% identity among species of the particular bioagent of interest, of upstream and downstream nucleotide sequences which serve as sequence amplification primer sites; b) an intervening variable region which exhibits no greater than about 5% identity among species; and c) a separation of between about 30 and 1000 nucleotides, or no more than about 50-250 nucleotides, or no more than about 60-100 nucleotides, between the conserved regions.

As a non-limiting example, for identification of *Bacillus* species, the conserved sequence regions of the chosen bioagent identifying amplicon must be highly conserved among all *Bacillus* species while the variable region of the bioagent identifying amplicon is sufficiently variable such that the molecular masses of the amplification products of all species of *Bacillus* are distinguishable.

Bioagent identifying amplicons amenable to molecular mass determination are either of a length, size or mass compatible with the particular mode of molecular mass determination or compatible with a means of providing a predictable fragmentation pattern in order to obtain predictable fragments of a length compatible with the particular mode of molecular mass determination. Such means of providing a predictable fragmentation pattern of an amplification product include, but are not limited to, cleavage with restriction enzymes or cleavage primers, for example.

Identification of bioagents can be accomplished at different levels using intelligent primers suited to resolution of each individual level of identification. "Broad range survey" intelligent primers are designed with the objective of identifying a bioagent as a member of a particular division of bioagents. A "bioagent division" is defined as group of bioagents above the species level and includes but is not limited to: orders, families, classes, clades, genera or other such groupings of bioagents above the species level. As a non-limiting example, members of the *Bacillus/Clostridia* group or gamma-proteobacteria group may be identified as such by employing broad range survey intelligent primers such as primers which target 16S or 23S ribosomal RNA.

In some embodiments, broad range survey intelligent primers are capable of identification of bioagents at the species level. One main advantage of the detection methods of the present invention is that the broad range survey intelligent primers need not be specific for a particular bacterial species, or even genus, such as *Bacillus* or *Streptomyces*. Instead, the primers recognize highly conserved regions across hundreds of bacterial species including, but not limited to, the species described herein. Thus, the same broad range survey intelligent primer pair can be used to identify any desired bacterium because it will bind to the conserved regions that flank a variable region specific to a single species, or common to several bacterial species, allowing unbiased nucleic acid amplification of the intervening sequence and determination of its molecular weight and base composition. For example, the 16S_971-1062, 16S_1228-1310 and 16S_1100-1188 regions are 98-99% conserved in about 900 species of bacteria (16S=16S rRNA, numbers indicate nucleotide position). In one embodiment of the present invention, primers used in the present method bind to one or more of these regions or portions thereof.

Due to their overall conservation, the flanking rRNA primer sequences serve as good intelligent primer binding sites to amplify the nucleic acid region of interest for most, if not all, bacterial species. The intervening region between the sets of primers varies in length and/or composition, and thus provides a unique base composition signature. Examples of intelligent primers that amplify regions of the 16S and 23S rRNA are shown in Figures 1A-1H. A typical primer amplified region in 16S rRNA is shown in Figure 2. The arrows represent primers that bind to highly conserved regions which flank a variable region in 16S rRNA domain III. The amplified region is the stem-loop structure under "1100-1188." It is advantageous to design the broad range survey intelligent primers to minimize the number of primers required for the analysis, and to allow detection of multiple members of a bioagent division using a single pair of primers. The advantage of using broad range survey intelligent primers is that once a bioagent is broadly identified, the process of further identification at species and sub-species levels is facilitated by directing the choice of additional intelligent primers.

"Division-wide" intelligent primers are designed with an objective of identifying a bioagent at the species level. As a non-limiting example, a *Bacillus anthracis, Bacillus cereus* and *Bacillus thuringiensis* can be distinguished from each other using division-wide intelligent primers. Division-wide intelligent primers are not always required for identification at the species level because broad range survey intelligent primers may provide sufficient identification resolution to accomplishing this identification objective.

"Drill-down" intelligent primers are designed with an objective of identifying a sub-species characteristic of a bioagent. A "sub-species characteristic" is defined as a property imparted to a bioagent at the sub-species level of identification as a result of the presence or absence of a particular segment of nucleic acid. Such sub-species characteristics include, but are not limited to, strains, sub-types, pathogenicity markers such as antibiotic resistance genes, pathogenicity islands, toxin genes and virulence factors. Identification of such sub-species characteristics is often critical for determining proper clinical treatment of pathogen infections.

Chemical Modifications of Intelligent Primers

Ideally, intelligent primer hybridization sites are highly conserved in order to facilitate the hybridization of the primer. In cases where primer hybridization is less efficient due to lower levels of conservation of sequence, intelligent primers can be chemically modified to improve the efficiency of hybridization.

For example, because any variation (due to codon wobble in the 3rd position) in these conserved regions among species is likely to occur in the third position of a DNA triplet, oligonucleotide primers can be designed such that the nucleotide corresponding to this position is a base which can bind to more than one nucleotide, referred to herein as a "universal base." For example, under this "wobble" pairing, inosine (I) binds to U, C or A;

guanine (G) binds to U or C, and uridine (U) binds to U or C. Other examples of universal bases include nitroindoles such as 5-nitroindole or 3-nitropyrrole (Loakes *et al.*, *Nucleosides and Nucleotides*, **1995**, *14*, 1001-1003), the degenerate nucleotides d**P** or d**K** (Hill *et al.*), an acyclic nucleoside analog containing 5-nitroindazole (Van Aerschot *et al.*, *Nucleosides and Nucleotides*, **1995**, *14*, 1053-1056) or the purine analog 1-(2-deoxy-β-D-ribofuranosyl)-imidazole-4-carboxamide (Sala *et al.*, *Nucl. Acids Res.*, **1996**, *24*, 3302-3306).

In another embodiment of the invention, to compensate for the somewhat weaker binding by the "wobble" base, the oligonucleotide primers are designed such that the first and second positions of each triplet are occupied by nucleotide analogs which bind with greater affinity than the unmodified nucleotide. Examples of these analogs include, but are not limited to, 2,6-diaminopurine which binds to thymine, propyne T which binds to adenine and propyne C and phenoxazines, including G-clamp, which binds to G. Propynylated pyrimidines are described in U.S. Patent Nos. 5,645,985, 5,830,653 and 5,484,908, each of which is commonly owned and incorporated herein by reference in its entirety. Propynylated primers are claimed in U.S Serial No. 10/294,203 which is also commonly owned and incorporated herein by reference in entirety. Phenoxazines are described in U.S. Patent Nos. 5,502,177, 5,763,588, and 6,005,096, each of which is incorporated herein by reference in its entirety. G-clamps are described in U.S. Patent Nos. 6,007,992 and 6,028,183, each of which is incorporated herein by reference in its entirety.

D. Characterization of Bioagent Identifying Amplicons

A theoretically ideal bioagent detector would identify, quantify, and report the complete nucleic acid sequence of every bioagent that reached the sensor. The complete sequence of the nucleic acid component of a pathogen would provide all relevant information about the threat, including its identity and the presence of drug-resistance or pathogenicity markers. This ideal has not yet been achieved. However, the present invention provides a straightforward strategy for obtaining information with the same practical value based on analysis of bioagent identifying amplicons by molecular mass determination.

In some cases, a molecular mass of a given bioagent identifying amplicon alone does not provide enough resolution to unambiguously identify a given bioagent. For example, the molecular mass of the bioagent identifying amplicon obtained using the intelligent primer pair "16S_971" would be 55622 Da for both *E. coli* and *Salmonella typhimurium*. However, if additional intelligent primers are employed to analyze additional bioagent identifying amplicons, a "triangulation identification" process is enabled. For example, the "16S_1100" intelligent primer pair yields molecular masses of 55009 and 55005 Da for *E. coli* and *Salmonella typhimurium,* respectively. Furthermore, the "23S_855" intelligent primer pair yields molecular masses of 42656 and 42698 Da for *E. coli* and *Salmonella typhimurium,* respectively. In this basic example, the second and third intelligent primer pairs provided the additional "fingerprinting" capability or resolution to distinguish between the two bioagents.

In another embodiment, the triangulation identification process is pursued by measuring signals from a plurality of bioagent identifying amplicons selected within multiple core genes. This process is used to reduce false negative and false positive signals, and enable reconstruction of the origin of hybrid or otherwise engineered bioagents. In this process, after identification of multiple core genes, alignments are created from nucleic acid sequence databases. The alignments are then analyzed for regions of conservation and variation, and bioagent identifying amplicons are selected to distinguish bioagents based on specific genomic differences. For example, identification of the three part toxin genes typical of *B. anthracis* (Bowen *et al., J. Appl. Microbiol.*, **1999,** *87,* 270-278) in the absence of the expected signatures from the *B. anthracis* genome would suggest a genetic engineering event.

The triangulation identification process can be pursued by characterization of bioagent identifying amplicons in a massively parallel fashion using the polymerase chain reaction (PCR), such as multiplex PCR, and mass spectrometric (MS) methods. Sufficient quantities of nucleic acids should be present for detection of bioagents by MS. A wide variety of techniques for preparing large amounts of purified nucleic acids or fragments thereof are well known to those of skill in the art. PCR requires one or more pairs of oligonucleotide primers that bind to regions which flank the target sequence(s) to be amplified. These primers prime synthesis of a different strand of DNA, with synthesis occurring in the direction of one primer towards the other primer. The primers, DNA to be amplified, a thermostable DNA polymerase (e.g. *Taq* polymerase), the four deoxynucleotide triphosphates, and a buffer are combined to initiate DNA synthesis. The solution is denatured by heating, then cooled to allow annealing of newly added primer, followed by another round of DNA synthesis. This process is typically repeated for about 30 cycles, resulting in amplification of the target sequence.

Although the use of PCR is suitable, other nucleic acid amplification techniques may also be used, including ligase chain reaction (LCR) and strand displacement amplification

(SDA). The high-resolution MS technique allows separation of bioagent spectral lines from background spectral lines in highly cluttered environments.

In another embodiment, the detection scheme for the PCR products generated from the bioagent(s) incorporates at least three features. First, the technique simultaneously detects and differentiates multiple (generally about 6-10) PCR products. Second, the technique provides a molecular mass that uniquely identifies the bioagent from the possible primer sites. Finally, the detection technique is rapid, allowing multiple PCR reactions to be run in parallel.

E. Mass Spectrometric Characterization of Bioagent Identifying Amplicons

Mass spectrometry (MS)-based detection of PCR products provides a means for determination of BCS which has several advantages. MS is intrinsically a parallel detection scheme without the need for radioactive or fluorescent labels, since every amplification product is identified by its molecular mass. The current state of the art in mass spectrometry is such that less than femtomole quantities of material can be readily analyzed to afford information about the molecular contents of the sample. An accurate assessment of the molecular mass of the material can be quickly obtained, irrespective of whether the molecular weight of the sample is several hundred, or in excess of one hundred thousand atomic mass units (amu) or Daltons. Intact molecular ions can be generated from amplification products using one of a variety of ionization techniques to convert the sample to gas phase. These ionization methods include, but are not limited to, electrospray ionization (ES), matrix-assisted laser desorption ionization (MALDI) and fast atom bombardment (FAB). For example, MALDI of nucleic acids, along with examples of matrices for use in MALDI of nucleic acids, are described in WO 98/54751 (Genetrace, Inc.).

In some embodiments, large DNAs and RNAs, or large amplification products therefrom, can be digested with restriction endonucleases prior to ionization. Thus, for example, an amplification product that was 10 kDa could be digested with a series of restriction endonucleases to produce a panel of, for example, 100 Da fragments. Restriction endonucleases and their sites of action are well known to the skilled artisan. In this manner, mass spectrometry can be performed for the purposes of restriction mapping.

Upon ionization, several peaks are observed from one sample due to the formation of ions with different charges. Averaging the multiple readings of molecular mass obtained from a single mass spectrum affords an estimate of molecular mass of the bioagent.

Electrospray ionization mass spectrometry (ESI-MS) is particularly useful for very high molecular weight polymers such as proteins and nucleic acids having molecular weights greater than 10 kDa, since it yields a distribution of multiply-charged molecules of the sample without causing a significant amount of fragmentation.

The mass detectors used in the methods of the present invention include, but are not limited to, Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR-MS), ion trap, quadrupole, magnetic sector, time of flight (TOF), Q-TOF, and triple quadrupole.

In general, the mass spectrometric techniques which can be used in the present invention include, but are not limited to, tandem mass spectrometry, infrared multiphoton dissociation and pyrolytic gas chromatography mass spectrometry (PGC-MS). In one embodiment of the invention, the bioagent detection system operates continually in bioagent detection mode using pyrolytic GC-MS without PCR for rapid detection of increases in biomass (for example, increases in fecal contamination of drinking water or of germ warfare agents). To achieve minimal latency, a continuous sample stream flows directly into the PGC-MS combustion chamber. When an increase in biomass is detected, a PCR process is automatically initiated. Bioagent presence produces elevated levels of large molecular fragments from, for example, about 100-7,000 Da which are observed in the PGC-MS spectrum. The observed mass spectrum is compared to a threshold level and when levels of biomass are determined to exceed a predetermined threshold, the bioagent classification process described hereinabove (combining PCR and MS, such as FT-ICR MS) is initiated. Optionally, alarms or other processes (halting ventilation flow, physical isolation) are also initiated by this detected biomass level.

The accurate measurement of molecular mass for large DNAs is limited by the adduction of cations from the PCR reaction to each strand, resolution of the isotopic peaks from natural abundance ^{13}C and ^{15}N isotopes, and assignment of the charge state for any ion. The cations are removed by in-line dialysis using a flow-through chip that brings the solution containing the PCR products into contact with a solution containing ammonium acetate in the presence of an electric field gradient orthogonal to the flow. The latter two problems are addressed by operating with a resolving power of >100,000 and by incorporating isotopically depleted nucleotide triphosphates into the DNA. The resolving power of the instrument is also a consideration. At a resolving power of 10,000, the modeled signal from the $[M-14H+]^{14-}$ charge state of an 84mer PCR product is poorly characterized and assignment of the charge state or exact mass is impossible. At a resolving power of 33,000, the peaks from the individual isotopic components are visible. At a resolving power of 100,000, the isotopic peaks are resolved to the baseline and assignment of the charge state for the ion is straightforward. The [^{13}C,^{15}N]-depleted triphosphates are obtained, for example, by growing microorganisms on depleted media and harvesting the nucleotides (Batey *et al.*, *Nucl. Acids Res.*, **1992**, *20*, 4515-4523).

While mass measurements of intact nucleic acid regions are believed to be adequate to determine most bioagents, tandem mass spectrometry (MS^n) techniques may provide more definitive information pertaining to molecular identity or sequence. Tandem MS involves the coupled use of two or more stages of mass analysis where both the separation and detection steps are based on mass spectrometry. The first stage is used to select an ion or component of a sample from which further structural information is to be obtained. The selected ion is then fragmented using, e.g., blackbody irradiation, infrared multiphoton dissociation, or collisional activation. For example, ions generated by electrospray ionization (ESI) can be fragmented using IR multiphoton dissociation. This activation leads to dissociation of glycosidic bonds and the phosphate backbone, producing two series of fragment ions, called the *w*-series (having an intact 3' terminus and a 5' phosphate following internal cleavage) and the *a*-Base series(having an intact 5' terminus and a 3' furan).

The second stage of mass analysis is then used to detect and measure the mass of these resulting fragments of product ions. Such ion selection followed by fragmentation routines can be performed multiple times so as to essentially completely dissect the molecular sequence of a sample.

If there are two or more targets of similar molecular mass, or if a single amplification reaction results in a product which has the same mass as two or more bioagent reference standards, they can be distinguished by using mass-modifying "tags." In this embodiment of the invention, a nucleotide analog or "tag" is incorporated during amplification (e.g., a 5-(trifluoromethyl) deoxythymidine triphosphate) which has a different molecular weight than the unmodified base so as to improve distinction of masses. Such tags are described in, for example, PCT WO97/33000, which is incorporated herein by reference in its entirety. This further limits the number of possible base compositions consistent with any mass. For example, 5-(trifluoromethyl)deoxythymidine triphosphate can be used in place of dTTP in a separate nucleic acid amplification reaction. Measurement of the mass shift between a conventional amplification product and the tagged product is used to quantitate the number of thymidine nucleotides in each of the single strands. Because the strands are complementary, the number of adenosine nucleotides in each strand is also determined.

In another amplification reaction, the number of G and C residues in each strand is determined using, for example, the cytidine analog 5-methylcytosine (5-meC) or propyne C. The combination of the A/T reaction and G/C reaction, followed by molecular weight determination, provides a unique base composition. This method is summarized in Figure 4 and Table 1.

Table 1

Mass tag	**Double strand sequence**	**Single strand Sequence**	**Total mass this strand**	**Base info this strand**	**Base info other strand**	**Total base comp. Top strand**	**Total base comp. Bottom strand**
T*mass (T*-T) = x	T*ACGT*ACGT* AT*GCAT*GCA	T*ACGT*ACGT*	3x	3T	3A	3T 2A 2C 2G	3A 2T 2G 2C
		AT*GCAT*GCA	2x	2T	2A		
C*mass (C*-C) = y	TAC*GTAC*GT ATGC*ATGC*A	TAC*GTAC*GT	2x	2C	2G		
		ATGC*ATGC*A	2x	2C	2G		

The mass tag phosphorothioate A (A*) was used to distinguish a *Bacillus anthracis* cluster. The *B. anthracis* ($A_{14}G_9C_{14}T_9$) had an average MW of 14072.26, and the *B. anthracis* ($A_1A^*_{13}G_9C_{14}T_9$) had an average molecular weight of 14281.11 and the phosphorothioate A had an average molecular weight of +16.06 as determined by ESI-TOF MS. The deconvoluted spectra are shown in Figure 5.

In another example, assume the measured molecular masses of each strand are 30,000.115Da and 31,000.115 Da respectively, and the measured number of dT and dA residues are (30,28) and (28,30). If the molecular mass is accurate to 100 ppm, there are 7 possible combinations of dG+dC possible for each strand. However, if the measured molecular mass is accurate to 10 ppm, there are only 2 combinations of dG+dC, and at 1 ppm accuracy there is only one possible base composition for each strand.

Signals from the mass spectrometer may be input to a maximum-likelihood detection and classification algorithm such as is widely used in radar signal processing. The detection processing uses matched filtering of BCS observed in mass-basecount space and allows for detection and subtraction of signatures from known, harmless organisms, and for detection of unknown bioagent threats. Comparison of newly observed bioagents to known bioagents is also possible, for estimation of threat level, by comparing their BCS to those of known organisms and to known forms of pathogenicity enhancement, such as insertion of antibiotic resistance genes or toxin genes.

Processing may end with a Bayesian classifier using log likelihood ratios developed from the observed signals and average background levels. The program emphasizes performance predictions culminating in probability-of-detection versus probability-of-false-alarm plots for conditions involving complex backgrounds of naturally occurring organisms and environmental contaminants. Matched filters consist of a priori expectations of signal values given the set of primers used for each of the bioagents. A genomic sequence database (e.g. GenBank) is used to define the mass basecount matched filters. The database contains known threat agents and benign background organisms. The latter is used to estimate and subtract the signature produced by the background organisms. A maximum likelihood detection of known background organisms is implemented using matched filters and a running-sum estimate of the noise covariance. Background signal strengths are estimated and used along with the matched filters to form signatures which are then subtracted. the maximum likelihood process is applied to this "cleaned up" data in a similar manner employing matched filters for the organisms and a running-sum estimate of the noise-covariance for the cleaned up data.

F. Base Composition Signatures as Indices of Bioagent Identifying Amplicons

Although the molecular mass of amplification products obtained using intelligent primers provides a means for identification of bioagents, conversion of molecular mass data to a base composition signature is useful for certain analyses. As used herein, a "base composition signature" (BCS) is the exact base composition determined from the molecular mass of a bioagent identifying amplicon. In one embodiment, a BCS provides an index of a specific gene in a specific organism.

Base compositions, like sequences, vary slightly from isolate to isolate within species. It is possible to manage this diversity by building "base composition probability clouds" around the composition constraints for each species. This permits identification of organisms in a fashion similar to sequence analysis. A "pseudo four-dimensional plot" can be used to visualize the concept of base composition probability clouds (Figure 18). Optimal primer design requires optimal choice of bioagent identifying amplicons and maximizes the separation between the base composition signatures of individual bioagents. Areas where clouds overlap indicate regions that may result in a misclassification, a problem which is overcome by selecting primers that provide information from different bioagent identifying amplicons, ideally maximizing the separation of base compositions. Thus, one aspect of the utility of an analysis of base composition probability clouds is that it provides a means for screening primer sets in order to avoid potential misclassifications of BCS and bioagent identity. Another aspect of the utility of base composition probability clouds is that they provide a means for predicting the identity of a bioagent whose exact measured BCS was not previously observed and/or indexed in a BCS database due to evolutionary transitions in its nucleic acid sequence.

It is important to note that, in contrast to probe-based techniques, mass spectrometry determination of base composition does not require prior knowledge of the composition in order to make the measurement, only to interpret the results. In this regard, the present invention provides bioagent classifying information similar to DNA sequencing and phylogenetic analysis at a level sufficient to detect and identify a given bioagent. Furthermore, the process of determination of a previously unknown BCS for a given bioagent (for example, in a case where sequence information is unavailable) has downstream utility by providing additional bioagent indexing information with which to populate BCS databases. The process of future bioagent identification is thus greatly improved as more BCS indexes become available in the BCS databases.

Another embodiment of the present invention is a method of surveying bioagent samples that enables detection and identification of all bacteria for which sequence information is available using a set of twelve broad-range intelligent PCR primers. Six of the twelve primers are "broad range survey primers" herein defined as primers targeted to broad divisions of bacteria (for example, the *Bacillus/Clostridia* group or gamma-proteobacteria). The other six primers of the group of twelve primers are "division-wide" primers herein defined as primers which provide more focused coverage and higher resolution. This method enables identification of nearly 100% of known bacteria at the species level. A further example of this embodiment of the present invention is a method herein designated "survey/drill-down" wherein a subspecies characteristic for detected bioagents is obtained using additional primers. Examples of such a subspecies characteristic include but are not limited to: antibiotic resistance, pathogenicity island, virulence factor, strain type, sub-species type, and clade group. Using the survey/drill-down method, bioagent detection, confirmation and a subspecies characteristic can be provided within hours. Moreover, the survey/drill-down method can be focused to identify bioengineering events such as the insertion of a toxin gene into a bacterial species that does not normally make the toxin.

G. Fields of Application of the Present Invention

The present methods allow extremely rapid and accurate detection and identification of bioagents compared to existing methods. Furthermore, this rapid detection and identification is possible even when sample material is impure. The methods leverage ongoing biomedical research in virulence, pathogenicity, drug resistance and genome sequencing into a method which provides greatly improved sensitivity, specificity and reliability compared to existing methods, with lower rates of false positives. Thus, the methods are useful in a wide variety of fields, including, but not limited to, those fields discussed below.

1. Forensic Investigations of Biowarfare Agents

In other embodiments of the invention, the methods disclosed herein can be used for epidemiological and forensics investigations. As used herein, "epidemiology" refers to an investigative process which attempts to link the effects of exposure to harmful agents to disease or mortality. As used herein, "forensics" is the study of evidence discovered at a crime investigation or accident scene and which may be used in a court of law. "Forensic science" is any science used for the purposes of the law, and therefore provides impartial scientific evidence for use in the courts of law, and in a criminal investigation and trial. Forensic science is a multidisciplinary subject, drawing principally from chemistry and biology, but also from physics, geology, psychology and social science, for example.

Epidemiological and forensic investigations of biowarfare- or bioterrorism-associated events can benefit from rapid and reliable methods of genetic analysis capable of characterizing a variety of genetic marker types. Examples of such genetic marker analyses include, but are not limited to: Multiple-Locus VNTR Analysis (MLVA), Multi-Locus Sequence Typing (MLST) and Single Nucleotide Polymorphism (SNP) analysis. These methods traditionally require independent PCR-based assays followed by electrophoresis on fluorescent sequencing platforms.

In addition, epidemiologists, for example, can use the present methods to determine the geographic origin of a particular strain of a protist or fungus. For example, a particular strain of bacteria or virus may have a sequence difference that is associated with a particular area of a country or the world and identification of such a sequence difference can lead to the identification of the geographic origin and epidemiological tracking of the spread of the particular disease, disorder or condition associated with the detected protist or fungus. In addition, carriers of particular DNA or diseases, such as mammals, non-mammals, birds, insects, and plants, can be tracked by screening their mtDNA. Diseases, such as malaria, can be tracked by screening the mtDNA of commensals such as mosquitoes.

In one embodiment the methods of the present invention are employed for identification of bioagent associated with an act of biowarfare, terrorism or criminal activity.

Examples of bioagents that may be used in acts of biowarfare, or bioterrorism include, but are not limited to: viruses such as: Crimean-Congo hemorrhagic fever virus, Eastern Equine Encephalitis virus, Ebola virus, Equine Morbillivirus, Flexal virus, Guanarito virus, Hantaan or other Hanta viruses, Junin virus, Lassa fever virus, Machupo virus, Marburg virus, Omsk hemorrhagic fever virus, Rift Valley fever virus, Russian Spring-Summer encephalitis virus, Sabia virus, Tick-borne encephalitis complex viruses, Variola major virus (Smallpox virus), Venezuelan Equine Encephalitis virus, Coronavirus and Yellow fever virus; bacteria such as: *Bacillus anthracis, Brucella abortus, Brucella melitensis, Brucella suis, Burkholderia* (*Pseudomonas*) *mallei, Burkholderia* (*Pseudomonas*) *pseudomallei, Clostridium botulinum, Francisella tularensis, Yersinia pestis, Rickettsiae, Coxiella burnetii, Rickettsia prowazekii,* and *Rickettsia rickettsii*; fungi such as: *Coccidioides immitis*; and toxins such as: abrin, aflatoxins, *Botulinum* toxins, *Clostridium perfringens* epsilon toxin, conotoxins, diacetoxyscirpenol, ricin, saxitoxin, shigatoxin, *Staphylococcal* enterotoxins, Tetrodotoxin and T-2 toxin (see, for example, www.ehs.iastate.edu/publications/factsheets/bioterrorismlaws.pdf).

As an example, a bioagent such as *Bacillus anthracis* is identified in a sample obtained from the site of an incident of biowarfare, terrorism or crime by employing intelligent primers to amplify a bioagent identifying amplicon from the bioagent, determining molecular mass (and the base composition signature if desired) of the amplified nucleic acid and matching the molecular mass (and base composition signature, if desired) with a molecular mass or base composition indexed to a bioagent contained in a database of reference base composition signatures.

2. Epidemiologic Investigations by Genotyping of Bioagents

In another embodiment, a bioagent is genotyped. "Genotyping" as used herein, refers to characterization of a nucleic acid representing a gene or a portion of a gene. As a nonlimiting example, genotyping is carried out to investigate the presence or absence of "pathogenicity factors" defined herein as a segment of nucleic acid which confers pathogenic properties upon a bioagent. Examples of pathogenicity factors include, but are not limited to: pathogenicity islands, virulence markers and toxin components such as: protective antigen, lethal factor and edema factor, all of which are found on the plasmid pX01 and the antiphagocytic capsule found on plasmid pX02 of *Bacillus anthracis*. Primers targeting nucleic acid regions suspected of containing pathogenicity factors are used to amplify the nucleic acid whose, molecular mass (and base composition signature if desired) is then determined and compared to molecular masses (and base composition signatures if desired) of known pathogenicity factors to identify the pathogenicity factor.

An example of an advantage conferred by genotyping is that genetic engineering events are detected. Genetic engineering has been used in the past decade to alter the genes of biological weapon agents. Researchers in the USA, UK, Russia, Germany and other countries have introduced genes into hazardous bacteria that are likely to enhance the biowarfare possibilities of these microbes. Strains have been designed that can withstand antibiotics, are undetectable by traditional equipment, can overcome vaccines, or that cause unusual symptoms, thereby hampering diagnosis. In general, gene transfer is used to build more effective biological weapons, it could be used to broaden the military biological warfare spectrum, making it more difficult to fight and control biowarfare agents (www.sunshine-project.org/publications/pr/pr130700.html).

The present invention is particularly well suited for the task of detecting genetic engineering events since reference databases can be populated with molecular mass data or associated base composition signatures for known pathogenic factors. Primers targeting nucleic acid regions suspected of containing pathogenicity factors are used to amplify the nucleic acid whose molecular mass and base composition signature is then determined and compared to molecular masses and base composition signatures of known pathogenicity factors. A match between a molecular mass and/or base composition signature determined for genetically engineered bioagent and the base composition of a pathogenicity factor provides a means for identifying the pathogenicity factor incorporated into the genetically engineered bioagent so that appropriate countermeasures may be carried out.

In another embodiment, the present methods are used in forensic investigations which employ microbial geographic profiling information to track a known or suspected terrorist or criminal by obtaining bioagent samples from the site of incidence of an act of terrorism or crime, identifying the bioagent and correlating the identity of the bioagent with the likelihood that the bioagent is associated with the known or suspected terrorist or criminal. As used herein, "microbial geographic profiling" refers to the process of associating a particular genetic characteristic of a microbial bioagent with the geographic location in which it originates and is typically located.

In some embodiments of the invention the travels of a known terrorist are tracked. The terrorist can be a member of a known terrorist organization, such as Al Qaeda, or can be a member of an unknown terrorist organization. Alternately, the terrorist can be a single entity operating independently of any organization. Known terrorists include those individuals who are known by or listed by particular governments, such as the United States, or departments or agencies within a government, such as the Federal Bureau of Investigation, the Central Intelligence Agency, the Department of Homeland Security, the State Department, Congress, the National Security Agency, and the like. Suspected terrorists include those individuals not known to be terrorists as described above, but who are likely to be or are suspected of supporting or engaging in terroristic acts.

Travels of a known or suspected terrorist include all geographic movements of a particular individual terrorist. The travels of the individual terrorist may represent geographic movements of a terrorist organization of which the individual terrorist is a member. Geographic movements or travels of a terrorist include, foot travel, air travel, automobile travel, train travel, bus travel, or any other means of movement from one location in the world to another location in the world. Tracking of such travels or geographic movements includes determining at least one geographic location that the terrorist has occupied, other than the geographic location the terrorist is in when a sample is taken from the terrorist. For example, the travels of a terrorist that is currently located, for instance, in the United Kingdom, at the time a sample is obtained can be "tracked" to another location such as, for instance, Afghanistan. Thus, the travels of such a terrorist can be tracked from Afghanistan to the United Kingdom.

Geographic locations of interest to be tracked include every country or state in the world including, but not limited to, the Mideast (e.g., Afghanistan, Iraq, Iran, Syria, Jordan, Palestinian-occupied territory, Lebanon, Kuwait, Yemen, United Arab Emirates, and Saudi Arabia), Northern Africa (e.g., Egypt, Sudan, Somalia, Tunisia, Morocco, and Libya), Asia (e.g., North Korea, China, Pakistan, India, Philippines, and Indonesia), as well as other countries, states, or territories known or suspected of sponsoring or harboring terrorists.

A nucleic acid from a bioagent obtained from a sample associated with the terrorist is obtained. Samples associated with terrorists can be obtained knowingly from the terrorist or can be obtained covertly from the terrorist. For example, the terrorist can be detained and the sample or samples obtained from the terrorist with the terrorist's knowledge. Alternately, the sample or samples can be obtained covertly from the terrorists without the terrorist's knowledge. The sample from the terrorist can be associated with the terrorist in a direct manner or indirect manner. For example, samples associated with a terrorist can be obtained directly from the terrorist. Such samples include, but are not limited to, a sample of bodily fluid of a sample of tissue from the terrorist. Examples of bodily fluids include, but are not limited to, blood, sweat, tears, urine, saliva, and the like. Examples of bodily tissues include, but are not limited to, hair, skin, bone, and the like. In addition, types of bodily tissues include products derived therefrom such as feces, mucous, and the like. Thus, a food product containing a bioagent that has been consumed by a terrorist can be detected in samples of feces or saliva. Other samples associated with a terrorist include, but are not limited to, a sample of clothing from the terrorist, a sample of environmental material from the terrorist, a sample from a pet travelling with the terrorist, or a sample from a luggage traveling with the terrorist. Examples of environmental material include, but are not limited to, dirt, water, plant material, animal material, and the like that may be present somewhere on the terrorist or with pets, luggage, or companions traveling therewith.

In one example of the present embodiment, a terrorist from a camp in the Sahara Desert commits an act of terrorism in a European country, leaving forensic evidence at the scene of the terrorist act. Samples of the forensic evidence are analyzed by the methods of the present invention. *Bacillus mojavensis* is identified in the sample by employing intelligent primers to amplify nucleic acid from the bacterium, determining the base composition signature of the amplified nucleic acid and matching the base composition with a base composition indexed to *Bacillus mojavensis* contained in a database of reference base composition signatures. This analysis indicates that *Bacillus mojavensis* is present in the soil sample associated with the terrorist. Further genotyping of the bioagent using primers targeting the pTA-like plasmid *rep* and *mob* genes then provides an indication that the strain of *Bacillus mojavensis* is *B. mojavensis* IM-E-3, a strain found in the Sahara Desert (Mason, M.P. *et al. FEMS Microbiol. Ecol.* **2002**, *42*, 235-241), thus indicating that the terrorist is associated with a soil sample originating from the Sahara Desert. These analyses are optionally repeated for additional bioagents identified in the forensic sample to ascertain other geographic locations to which the known or suspected terrorist has traveled.

While the present invention has been described with specificity in accordance with certain of its embodiments, the following examples serve only to illustrate the invention and are not intended to limit the same.

EXAMPLES

Example 1: Nucleic Acid Isolation and PCR

In one embodiment, nucleic acid is isolated from the organisms and amplified by PCR using standard methods prior to BCS determination by mass spectrometry. Nucleic acid is isolated, for example, by detergent lysis of bacterial cells, centrifugation and ethanol precipitation. Nucleic acid isolation methods are described in, for example, *Current Protocols in Molecular Biology* (Ausubel et al.) and *Molecular Cloning; A Laboratory Manual* (Sambrook *et al.*). The nucleic acid is then amplified using standard methodology, such as PCR, with primers which bind to conserved regions of the nucleic acid which contain an intervening variable sequence as described below.

General Genomic DNA Sample Prep Protocol: Raw samples are filtered using Supor-200 0.2 µm membrane syringe filters (VWR International) . Samples are transferred to 1.5 ml eppendorf tubes pre-filled with 0.45 g of 0.7 mm Zirconia beads followed by the addition of 350 µl of ATL buffer (Qiagen, Valencia, CA). The samples are subjected to bead beating for 10 minutes at a frequency of 19 1/s in a Retsch Vibration Mill (Retsch). After centrifugation, samples are transferred to an S-block plate (Qiagen) and DNA isolation is completed with a BioRobot 8000 nucleic acid isolation robot (Qiagen).

Swab Sample Protocol: Allegiance S/P brand culture swabs and collection/transport system are used to collect samples. After drying, swabs are placed in 17x100 mm culture tubes (VWR International) and the genomic nucleic acid isolation is carried out automatically with a Qiagen Mdx robot and the Qiagen QIAamp DNA Blood BioRobot Mdx genomic preparation kit (Qiagen, Valencia, CA).

Example 2: Mass spectrometry

FTICR Instrumentation: The FTICR instrument is based on a 7 tesla actively shielded superconducting magnet and modified Bruker Daltonics Apex II 70e ion optics and vacuum chamber. The spectrometer is interfaced to a LEAP PAL autosampler and a custom fluidics control system for high throughput screening applications. Samples are analyzed directly from 96-well or 384-well microtiter plates at a rate of about 1 sample/minute. The Bruker data-acquisition platform is supplemented with a lab-built ancillary NT datastation which controls the autosampler and contains an arbitrary waveform generator capable of generating complex rf-excite waveforms (frequency sweeps, filtered noise, stored waveform inverse Fourier transform (SWIFT), etc.) for sophisticated tandem MS experiments. For oligonucleotides in the 20-30-mer regime typical performance characteristics include mass resolving power in excess of 100,000 (FWHM), low ppm mass measurement errors, and an operable *m/z* range between 50 and 5000 *m/z.*

Modified ESI Source: In sample-limited analyses, analyte solutions are delivered at 150 nL/minute to a 30 mm i.d. fused-silica ESI emitter mounted on a 3-D micromanipulator. The ESI ion optics consists of a heated metal capillary, an rf-only hexapole, a skimmer cone, and an auxiliary gate electrode. The 6.2 cm rf-only hexapole is comprised of 1 mm diameter rods and is operated at a voltage of 380 Vpp at a frequency of 5 MHz. A lab-built electro-mechanical shutter can be employed to prevent the electrospray plume from entering the inlet capillary unless triggered to the "open" position via a TTL pulse from the data station. When in the "closed" position, a stable electrospray plume is maintained between the ESI emitter and the face of the shutter. The back face of the shutter arm contains an elastomeric seal that can be positioned to form a vacuum seal with the inlet capillary. When the seal is removed, a 1 mm gap between the shutter blade and the capillary inlet allows constant pressure in the external ion reservoir regardless of whether the shutter is in the open or closed position. When the shutter is triggered, a "time slice" of ions is allowed to enter the inlet capillary and is subsequently accumulated in the external ion reservoir. The rapid response time of the ion shutter (< 25 ms) provides reproducible, user defined intervals during which ions can be injected into and accumulated in the external ion reservoir.

Apparatus for Infrared Multiphoton Dissociation: A 25 watt CW CO_2 laser operating at 10.6 μm has been interfaced to the spectrometer to enable infrared multiphoton dissociation (IRMPD) for oligonucleotide sequencing and other tandem MS applications. An aluminum optical bench is positioned approximately 1.5 m from the actively shielded superconducting magnet such that the laser beam is aligned with the central axis of the magnet. Using standard IR-compatible mirrors and kinematic mirror mounts, the unfocused 3 mm laser beam is aligned to traverse directly through the 3.5 mm holes in the trapping electrodes of the FTICR trapped ion cell and longitudinally traverse the hexapole region of the external ion guide finally impinging on the skimmer cone. This scheme allows IRMPD to be conducted in an *m/z* selective manner in the trapped ion cell (e.g. following a SWIFT isolation of the species of interest), or in a broadband mode in the high pressure region of the external ion reservoir where collisions with neutral molecules stabilize IRMPD-generated metastable fragment ions resulting in increased fragment ion yield and sequence coverage.

Example 3 :Identification of Bioagents

Table 2 shows a small cross section of a database of calculated molecular masses for over 9 primer sets and approximately 30 organisms. The primer sets were derived from rRNA alignment. Examples of regions from rRNA consensus alignments are shown in Figures 1A-1C. Lines with arrows are examples of regions to which intelligent primer pairs for PCR are designed. The primer pairs are >95% conserved in the bacterial sequence database (currently over 10,000 organisms). The intervening regions are variable in length and/or composition, thus providing the base composition "signature" (BCS) for each organism. Primer pairs were chosen so the total length of the amplified region is less than about 80-90 nucleotides. The label for each primer pair represents the starting and ending base number of the amplified region on the consensus diagram.

Included in the short bacterial database cross-section in Table 2 are many well known pathogens/biowarfare agents (shown in bold/red typeface) such as *Bacillus anthracis* or *Yersinia pestis* as well as some of the bacterial organisms found commonly in the natural environment such as *Streptomyces*. Even closely related organisms can be distinguished from each other by the appropriate choice of primers. For instance, two low G+C organisms, *Bacillus anthracis* and *Staph aureus*, can be distinguished from each other by using the primer pair defined by 16S_1337 or 23S_855 (ΔM of 4 Da).

Table 2: Cross Section Of A Database Of Calculated Molecular Masses[1]

Primer Regions -->	16S_971	16S_1100	16S_1337	16S_1294	16S_1228	23S_1021	23S_855	23S_193	23S_115
Bug Name									
Acinetobacter calcoaceticus	55619.1	55004	28446.7	35854.9	51295.4	30299	42654	39557.5	54999
Bacillus anthracis	**55005**	**54388**	**28448**	**35238**	**51296**	**30295**	**42651**	**39560**	**56850**
Bacillus cereus	55622.1	54387.9	28447.6	35854.9	51296.4	30295	42651	39560.5	56850.3
Bordetella bronchiseptica	56857.3	51300.4	28446.7	35857.9	51307.4	30299	42653	39559.5	51920.5
Borrelia burgdorferi	56231.2	55621.1	28440.7	35852.9	51295.4	30297	42029.9	38941.4	52524.6
Brucella abortus	**58098**	**55011**	**28448**	**35854**	**50683**				
Campylobacter jejuni	58088.5	54386.9	29061.8	35856.9	50674.3	30294	42032.9	39558.5	45732.5
Chlamydia pnuemoniae	**55000**	**55007**	**29063**	**35855**	**50676**	**30295**	**42036**	**38941**	**56230**
Clostridium botulinum	**55006**	**53767**	**28445**	**35855**	**51291**	**30300**	**42656**	**39562**	**54999**
Clostridium difficile	56855.3	54386.9	28444.7	35853.9	51296.4	30294	41417.8	39556.5	55612.2
Enterococcus faecalis	55620.1	54387.9	28447.6	35858.9	51296.4	30297	42652	39559.5	56849.3
Escherichia coli	**55622**	**55009**	**28445**	**35857**	**51301**	**30301**	**42656**	**39562**	**54999**
Francisella tularensis	**53769**	**54385**	**28445**	**35856**	**51298**				
Haemophilus influenzae	55620.1	55006	28444.7	35855.9	51298.4	30298	42656	39560.5	55613.1
Klebsiella pneumoniae	55622.1	55008	28442.7	35856.9	51297.4	30300	42655	39562.5	55000
Legionella pneumophila	**55618**	**55626**	**28446**	**35857**	**51303**				
Mycobacterium avium	54390.9	55631.1	29064.8	35858.9	51915.5	30298	42656	38942.4	56241.2
Mycobacterium leprae	54389.9	55629.1	29064.8	35860.9	51917.5	30298	42656	39559.5	56240.2
Mycobacterium tuberculosis	54390.9	55629.1	29064.8	35860.9	51301.4	30299	42656	39560.5	56243.2
Mycoplasma genitalium	53143.7	45115.4	29061.8	35854.9	50671.3	30294	43264.1	39558.5	56842.4
Mycoplasma pneumoniae	53143.7	45118.4	29061.8	35854.9	50673.3	30294	43264.1	39559.5	56843.4
Neisseria gonorrhoeae	55627.1	54389.9	28445.7	35855.9	51302.4	30300	42649	39561.5	55000
Pseudomonas aeruginosa	**55623**	**55010**	**28443**	**35858**	**51301**	**30298**	**43272**	**39558**	**55619**
Rickettsia prowazekii	**58093**	**55621**	**28448**	**35853**	**50677**	**30293**	**42650**	**39559**	**53139**
Rickettsia rickettsii	**58094**	**55623**	**28448**	**35853**	**50679**	**30293**	**42648**	**39559**	**53755**
Salmonella typhimurium	**55622**	**55005**	**28445**	**35857**	**51301**	**30301**	**42658**		
Shigella dysenteriae	**55623**	**55009**	**28444**	**35857**	**51301**				
Staphylococcus aureus	56854.3	54386.9	28443.7	35852.9	51294.4	30298	42655	39559.5	57466.4
Streptomyces	54389.9	59341.6	29063.8	35858.9	51300.4			39563.5	56864.3
Treponema pallidum	56245.2	55631.1	28445.7	35851.9	51297.4	30299	42034.9	38939.4	57473.4
Vibrio cholerae	**55625**	**55626**	**28443**	**35857**	**52536**	**29063**	**30303**	**35241**	**50675**
Vibrio parahaemolyticus	54384.9	55626.1	28444.7	34620.7	50064.2				
Yersinia pestis	**55620**	**55626**	**28443**	**35857**	**51299**				

[1] Molecular mass distribution of PCR amplified regions for a selection of organisms (rows) across various primer pairs (columns). Pathogens are shown in **bold.** Empty cells indicate presently incomplete or missing data.

Figure 6 shows the use of ESI-FT-ICR MS for measurement of exact mass. The spectra from 46mer PCR products originating at position 1337 of the 16S rRNA from *S. aureus* (upper) and *B. anthracis* (lower) are shown. These data are from the region of the spectrum containing signals from the $[M-8H+]^{8-}$ charge states of the respective 5'-3' strands. The two strands differ by two (AT→CG) substitutions, and have measured masses of 14206.396 and 14208.373 + 0.010 Da, respectively. The possible base compositions derived from the masses of the forward and reverse strands for the *B. anthracis* products are listed in Table 3.

Table 3: Possible base composition for *B. anthracis* products

Calc. Mass	Error	Base Comp.
14208.2935	0.079520	A1 G17 C10 T18
14208.3160	0.056980	A1 G20 C15 T10

14208.3386	0.034440	A1 G23 C20 T2
14208.3074	0.065560	A6 G11 C3 T26
14208.3300	0.043020	A6 G14 C8 T18
14208.3525	0.020480	A6 G17 C13 T10
14208.3751	0.002060	A6 G20 C18 T2
14208.3439	0.029060	A11 G8 C1 T26
14208.3665	0.006520	A11 G11 C6 T18
14208.3890	**0.016020**	**A11 G14 C11 T10**
14208.4116	0.038560	A11 G17 C16 T2
14208.4030	0.029980	A16 G8 C4 T18
14208.4255	0.052520	A16 G11 C9 T10
14208.4481	0.075060	A16 G14 C14 T2
14208.4395	0.066480	A21 G5 C2 T18
14208.4620	0.089020	A21 G8 C7 T10
14079.2624	0.080600	A0 G14 C13 T19
14079.2849	0.058060	A0 G17 C18 T11
14079.3075	0.035520	A0 G20 C23 T3
14079.2538	0.089180	A5 G5 C1 T35
14079.2764	0.066640	A5 G8 C6 T27
14079.2989	0.044100	A5 G11 C11 T19
14079.3214	0.021560	A5 G14 C16 T11
14079.3440	0.000980	A5 G17 C21 T3
14079.3129	0.030140	A10 G5 C4 T27
14079.3354	0.007600	A10 G8 C9 T19
14079.3579	**0.014940**	**A10 G11 C14 T11**
14079.3805	0.037480	A10 G14 C19 T3
14079.3494	0.006360	A15 G2 C2 T27
14079.3719	0.028900	A15 G5 C7 T19
14079.3944	0.051440	A15 G8 C12 T11
14079.4170	0.073980	A15 G11 C17 T3
14079.4084	0.065400	A20 G2 C5 T19
14079.4309	0.087940	A20 G5 C10 T13

Among the 16 compositions for the forward strand and the 18 compositions for the reverse strand that were calculated, only one pair (shown in **bold**) are complementary, corresponding to the actual base compositions of the *B. anthracis* PCR products.

Example 4: BCS of Region from *Bacillus anthracis* and *Bacillus cereus*

A conserved Bacillus region from *B. anthracis* ($A_{14}G_9C_{14}T_9$) and *B. cereus* ($A_{15}G_9C_{13}T_9$) having a C to A base change was synthesized and subjected to ESI-TOF MS. The results are shown in Figure 7 in which the two regions are clearly distinguished using the method of the present invention (MW=14072.26 vs. 14096.29).

Example 5: Identification of additional bioagents

In other examples of the present invention, the pathogen *Vibrio cholera* can be distinguished from *Vibrio parahemolyticus* with $\Delta M > 600$ Da using one of three 16S primer sets shown in Table 2 (16S_971, 16S_1228 or 16S_1294) as shown in Table 4. The two mycoplasma species in the list (*M. genitalium* and *M. pneumoniae*) can also be distinguished from each other, as can the three mycobacteriae. While the direct mass measurements of amplified products can identify and distinguish a large number of organisms, measurement of the base composition signature provides dramatically enhanced resolving power for closely related organisms. In cases such as *Bacillus anthracis* and *Bacillus cereus* that are virtually indistinguishable from each other based solely on mass differences, compositional analysis or fragmentation patterns are used to resolve the differences. The single base difference between the two organisms yields different fragmentation patterns, and despite the presence of the ambiguous/unidentified base N at position 20 in *B. anthracis*, the two organisms can be identified.

Tables 4a-b show examples of primer pairs from Table 1 which distinguish pathogens from background.

Table 4a

Organism name	23S_855	16S_1337	23S_1021
Bacillus anthracis	42650.98	28447.65	30294.98
Staphylococcus aureus	42654.97	28443.67	30297.96

Table 4b

Organism name	**16S_971**	**16S_1294**	**16S_1228**
Vibrio cholerae	55625.09	35856.87	52535.59
Vibrio parahaemolyticus	54384.91	34620.67	50064.19

Table 5 shows the expected molecular weight and base composition of region 16S_1100-1188 in *Mycobacterium avium* and *Streptomyces sp.*

Table 5

Region	**Organism name**	**Length**	**Molecular weight**	**Base comp.**
16S_1100-1188	*Mycobacterium avium*	82	25624.1728	$A_{16}G_{32}C_{18}T_{16}$
16S_1100-1188	*Streptomyces sp.*	96	29904.871	$A_{17}G_{38}C_{27}T_{14}$

Table 6 shows base composition (single strand) results for 16S_1100-1188 primer amplification reactions different species of bacteria. Species which are repeated in the table (e.g., *Clostridium botulinum*) are different strains which have different base compositions in the 16S_1100-1188 region.

Table 6

Organism name	**Base comp.**	**Organism name**	**Base comp.**
Mycobacterium avium	$A_{16}G_{32}C_{18}T_{16}$	*Vibrio cholerae*	$A_{23}G_{30}C_{21}T_{16}$
Streptomyces sp.	$A_{17}G_{38}C_{27}T_{14}$	***Aeromonas hydrophila***	**$A_{23}G_{31}C_{21}T_{15}$**
Ureaplasma urealyticum	$A_{18}G_{30}C_{17}T_{17}$	***Aeromonas salmonicida***	**$A_{23}G_{31}C_{21}T_{15}$**
Streptomyces sp.	$A_{19}G_{36}C_{24}T_{18}$	*Mycoplasma genitalium*	$A_{24}G_{19}C_{12}T_{18}$
Mycobacterium leprae	$A_{20}G_{32}C_{22}T_{16}$	*Clostridium botulinum*	$A_{24}G_{25}C_{18}T_{20}$
M. tuberculosis	**$A_{20}G_{33}C_{21}T_{16}$**	*Bordetella bronchiseptica*	$A_{24}G_{26}C_{19}T_{14}$
Nocardia asteroides	**$A_{20}G_{33}C_{21}T_{16}$**	*Francisella tularensis*	$A_{24}G_{26}C_{19}T_{19}$
Fusobacterium necroforum	$A_{21}G_{26}C_{22}T_{18}$	***Bacillus anthracis***	**$A_{24}G_{26}C_{20}T_{18}$**
Listeria monocytogenes	$A_{21}G_{27}C_{19}T_{19}$	***Campylobacter jejuni***	**$A_{24}G_{26}C_{20}T_{18}$**
Clostridium botulinum	$A_{21}G_{27}C_{19}T_{21}$	***Staphylococcus aureus***	**$A_{24}G_{26}C_{20}T_{18}$**
Neisseria gonorrhoeae	$A_{21}G_{28}C_{21}T_{18}$	*Helicobacter pylori*	$A_{24}G_{26}C_{20}T_{19}$
Bartonella quintana	$A_{21}G_{30}C_{22}T_{16}$	*Helicobacter pylori*	$A_{24}G_{26}C_{21}T_{18}$
Enterococcus faecalis	$A_{22}G_{27}C_{20}T_{19}$	*Moraxella catarrhalis*	$A_{24}G_{26}C_{23}T_{16}$

Bacillus megaterium	$A_{22}G_{28}C_{20}T_{18}$	*Haemophilus influenzae Rd*	$A_{24}G_{28}C_{20}T_{17}$
Bacillus subtilis	$A_{22}G_{28}C_{21}T_{17}$	***Chlamydia trachomatis***	$\mathbf{A_{24}G_{28}C_{21}T_{16}}$
Pseudomonas aeruginosa	$A_{22}G_{29}C_{23}T_{15}$	***Chlamydophila pneumoniae***	$\mathbf{A_{24}G_{28}C_{21}T_{16}}$
Legionella pneumophila	$A_{22}G_{32}C_{20}T_{16}$	***C. pneumonia AR39***	$\mathbf{A_{24}G_{28}C_{21}T_{16}}$
Mycoplasma pneumoniae	$A_{23}G_{20}C_{14}T_{16}$	*Pseudomonas putida*	$A_{24}G_{29}C_{21}T_{16}$
Clostridium botulinum	$A_{23}G_{26}C_{20}T_{19}$	***Proteus vulgaris***	$\mathbf{A_{24}G_{30}C_{21}T_{15}}$
Enterococcus faecium	$A_{23}G_{26}C_{21}T_{18}$	***Yersinia pestis***	$\mathbf{A_{24}G_{30}C_{21}T_{15}}$
Acinetobacter calcoaceti	$A_{23}G_{26}C_{21}T_{19}$	***Yersinia pseudotuberculos***	$\mathbf{A_{24}G_{30}C_{21}T_{15}}$
Leptospira borgpeterseni	$\mathbf{A_{23}G_{26}C_{24}T_{15}}$	*Clostridium botulinum*	$A_{25}G_{24}C_{18}T_{21}$
Leptospira interrogans	$\mathbf{A_{23}G_{26}C_{24}T_{15}}$	*Clostridium tetani*	$A_{25}G_{25}C_{18}T_{20}$
Clostridium perfringens	$A_{23}G_{27}C_{19}T_{19}$	*Francisella tularensis*	$A_{25}G_{25}C_{19}T_{19}$
Bacillus anthracis	$\mathbf{A_{23}G_{27}C_{20}T_{18}}$	*Acinetobacter calcoacetic*	$A_{25}G_{26}C_{20}T_{19}$
Bacillus cereus	$\mathbf{A_{23}G_{27}C_{20}T_{18}}$	*Bacteriodes fragilis*	$A_{25}G_{27}C_{16}T_{22}$
Bacillus thuringiensis	$\mathbf{A_{23}G_{27}C_{20}T_{18}}$	*Chlamydophila psittaci*	$A_{25}G_{27}C_{21}T_{16}$
Aeromonas hydrophila	$\mathit{A_{23}G_{29}C_{21}T_{16}}$	*Borrelia burgdorferi*	$A_{25}G_{29}C_{17}T_{19}$
Escherichia coli	$\mathit{A_{23}G_{29}C_{21}T_{16}}$	*Streptobacillus monilifor*	$A_{26}G_{26}C_{20}T_{16}$
Pseudomonas putida	$A_{23}G_{29}C_{21}T_{17}$	*Rickettsia prowazekii*	$A_{26}G_{28}C_{18}T_{18}$
Escherichia coli	$\mathbf{A_{23}G_{29}C_{22}T_{15}}$	*Rickettsia rickettsii*	$A_{26}G_{28}C_{20}T_{16}$
Shigella dysenteriae	$\mathbf{A_{23}G_{29}C_{22}T_{15}}$	*Mycoplasma mycoides*	$A_{28}G_{23}C_{16}T_{20}$

The same organism having different base compositions are different strains. Groups of organisms which are highlighted or in italics have the same base compositions in the amplified region. Some of these organisms can be distinguished using multiple primers. For example, *Bacillus anthracis* can be distinguished from *Bacillus cereus* and *Bacillus thuringiensis* using the primer 16S_971-1062 (Table 7). Other primer pairs which produce unique base composition signatures are shown in Table 6 (bold). Clusters containing very similar threat and ubiquitous non-threat organisms (e.g. *anthracis* cluster) are distinguished at high resolution with focused sets of primer pairs. The known biowarfare agents in Table 6 are *Bacillus anthracis, Yersinia pestis, Francisella tularensis* and *Rickettsia prowazekii.*

Table 7

Organism	**16S_971-1062**	**16S_1228-1310**	**16S_1100-1188**
Aeromonas hydrophila	$A_{21}G_{29}C_{22}T_{20}$	$A_{22}G_{27}C_{21}T_{13}$	$A_{23}G_{31}C_{21}T_{15}$
Aeromonas salmonicida	$A_{21}G_{29}C_{22}T_{20}$	$A_{22}G_{27}C_{21}T_{13}$	$A_{23}G_{31}C_{21}T_{15}$
Bacillus anthracis	$\mathbf{A_{21}G_{27}C_{22}T_{22}}$	$A_{24}G_{22}C_{19}T_{18}$	$A_{23}G_{27}C_{20}T_{18}$
Bacillus cereus	$A_{22}G_{27}C_{21}T_{22}$	$A_{24}G_{22}C_{19}T_{18}$	$A_{23}G_{27}C_{20}T_{18}$
Bacillus thuringiensis	$A_{22}G_{27}C_{21}T_{22}$	$A_{24}G_{22}C_{19}T_{18}$	$A_{23}G_{27}C_{20}T_{18}$
Chlamydia trachomatis	$\mathbf{A_{22}G_{26}C_{20}T_{23}}$	$\mathbf{A_{24}G_{23}C_{19}T_{16}}$	$A_{24}G_{28}C_{21}T_{16}$
Chlamydia pneumoniae AR39	$A_{26}G_{23}C_{20}T_{22}$	$A_{26}G_{22}C_{16}T_{18}$	$A_{24}G_{28}C_{21}T_{16}$
Leptospira borgpetersenii	$A_{22}G_{26}C_{20}T_{21}$	$A_{22}G_{25}C_{21}T_{15}$	$A_{23}G_{26}C_{24}T_{15}$
Leptospira interrogans	$A_{22}G_{26}C_{20}T_{21}$	$A_{22}G_{25}C_{21}T_{15}$	$A_{23}G_{26}C_{24}T_{15}$
Mycoplasma genitalium	$A_{28}G_{23}C_{15}T_{22}$	$\mathbf{A_{30}G_{18}C_{15}T_{19}}$	$\mathbf{A_{24}G_{19}C_{12}T_{18}}$
Mycoplasma pneumoniae	$A_{28}G_{23}C_{15}T_{22}$	$\mathbf{A_{27}G_{19}C_{16}T_{20}}$	$\mathbf{A_{23}G_{20}C_{14}T_{16}}$
Escherichia coli	$\mathbf{A_{22}G_{28}C_{20}T_{22}}$	$A_{24}G_{25}C_{21}T_{13}$	$A_{23}G_{29}C_{22}T_{15}$
Shigella dysenteriae	$\mathbf{A_{22}G_{28}C_{21}T_{21}}$	$A_{24}G_{25}C_{21}T_{13}$	$A_{23}G_{29}C_{22}T_{15}$
Proteus vulgaris	$\mathbf{A_{23}G_{26}C_{22}T_{21}}$	$\mathbf{A_{26}G_{24}C_{19}T_{14}}$	$A_{24}G_{30}C_{21}T_{15}$
Yersinia pestis	$A_{24}G_{25}C_{21}T_{22}$	$A_{25}G_{24}C_{20}T_{14}$	$A_{24}G_{30}C_{21}T_{15}$
Yersinia pseudotuberculosis	$A_{24}G_{25}C_{21}T_{22}$	$A_{25}G_{24}C_{20}T_{14}$	$A_{24}G_{30}C_{21}T_{15}$
Francisella tularensis	$\mathbf{A_{20}G_{25}C_{21}T_{23}}$	$\mathbf{A_{23}G_{26}C_{17}T_{17}}$	$\mathbf{A_{24}G_{26}C_{19}T_{19}}$
Rickettsia prowazekii	$\mathbf{A_{21}G_{26}C_{24}T_{25}}$	$\mathbf{A_{24}G_{23}C_{16}T_{19}}$	$\mathbf{A_{26}G_{28}C_{18}T_{18}}$
Rickettsia rickettsii	$\mathbf{A_{21}G_{26}C_{25}T_{24}}$	$\mathbf{A_{24}G_{24}C_{17}T_{17}}$	$\mathbf{A_{26}G_{28}C_{20}T_{16}}$

The sequence of *B. anthracis* and *B. cereus* in region 16S_971 is shown below. Shown in bold is the single base difference between the two species which can be detected using the methods of the present invention. *B. anthracis* has an ambiguous base at position 20.

*B.anthracis*_16S_971

GCGAAGAACCUUACCAGGU*N*UUGACAUCCUCUGACAACCCUAGAGAUAGGGCU UCUCCUUCGGGAGCAGAGUGACAGGUGGUGCAUGGUU (SEQ ID NO:1)

*B.cereus*_16S_971
GCGAAGAACCUUACCAGGUCUUGACAUCCUCUGAAAACCCUAGAGAUAGGGCU
UCUCCUUCGGGAGCAGAGUGACAGGUGGUGCAUGGUU (SEQ ID NO:2)

Example 6: ESI-TOF MS of sspE 56-mer Plus Calibrant

The mass measurement accuracy that can be obtained using an internal mass standard in the ESI-MS study of PCR products is shown in Fig.8. The mass standard was a 20-mer phosphorothioate oligonucleotide added to a solution containing a 56-mer PCR product from the *B. anthracis* spore coat protein sspE. The mass of the expected PCR product distinguishes *B. anthracis* from other species of Bacillus such as *B. thuringiensis* and *B. cereus*.

Example 7: *B. anthracis* ESI-TOF Synthetic 16S_1228 Duplex

An ESI-TOF MS spectrum was obtained from an aqueous solution containing 5 μM each of synthetic analogs of the expected forward and reverse PCR products from the nucleotide 1228 region of the *B. anthracis* 16S rRNA gene. The results (Fig. 9) show that the molecular weights of the forward and reverse strands can be accurately determined and easily distinguish the two strands. The $[M\text{-}21H^{+}]^{21-}$ and $[M\text{-}20H^{+}]^{20-}$ charge states are shown.

Example 8: ESI-FTICR-MS of Synthetic *B. anthracis* 16S_1337 46 Base Pair Duplex

An ESI-FTICR-MS spectrum was obtained from an aqueous solution containing 5 μM each of synthetic analogs of the expected forward and reverse PCR products from the nucleotide 1337 region of the *B. anthracis* 16S rRNA gene. The results (Fig. 10) show that the molecular weights of the strands can be distinguished by this method. The $[M\text{-}16H^{+}]^{16-}$ through $[M\text{-}10H^{+}]^{10-}$ charge states are shown. The insert highlights the resolution that can be realized on the FTICR-MS instrument, which allows the charge state of the ion to be determined from the mass difference between peaks differing by a single 13C substitution.

Example 9: ESI-TOF MS of 56-mer Oligonucleotide from saspB Gene of *B. anthracis* with Internal Mass Standard

ESI-TOF MS spectra were obtained on a synthetic 56-mer oligonucleotide (5 μM) from the saspB gene of *B. anthracis* containing an internal mass standard at an ESI of 1.7 μL/min as a function of sample consumption. The results (Fig. 11) show that the signal to noise is improved as more scans are summed, and that the standard and the product are visible after only 100 scans.

Example 10: ESI-TOF MS of an Internal Standard with Tributylammonium (TBA)-trifluoroacetate (TFA) Buffer

An ESI-TOF-MS spectrum of a 20-mer phosphorothioate mass standard was obtained following addition of 5 mM TBA-TFA buffer to the solution. This buffer strips charge from the oligonucleotide and shifts the most abundant charge state from $[M-8H^{+}]^{8-}$ to $[M-3H^{+}]^{3-}$ (Fig. 12).

Example 11: Master Database Comparison

The molecular masses obtained through Examples 1-10 are compared to molecular masses of known bioagents stored in a master database to obtain a high probability matching molecular mass.

Example 12: Master Data Base Interrogation over the Internet

The same procedure as in Example 11 is followed except that the local computer did not store the Master database. The Master database is interrogated over an internet connection, searching for a molecular mass match.

Example 13: Master Database Updating

The same procedure as in example 11 is followed except the local computer is connected to the internet and has the ability to store a master database locally. The local computer system periodically, or at the user's discretion, interrogates the Master database, synchronizing the local master database with the global Master database. This provides the current molecular mass information to both the local database as well as to the global Master database. This further provides more of a globalized knowledge base.

Example 14: Global Database Updating

The same procedure as in example 13 is followed except there are numerous such local stations throughout the world. The synchronization of each database adds to the diversity of information and diversity of the molecular masses of known bioagents.

Example 15: Genotyping of Diverse Strains of *Bacillis anthracis*

In accordance with the present invention, the present methods were employed for genotyping 24 globally diverse isolates of *B. anthracis* using both MLVA and SNP analyses.

For SNP analysis, bioagent identifying amplicons less than 100 bp long were designed around single or double SNPs from the protective antigen gene (PAG). Amplification products were obtained for the bioagent identifying amplicons. As shown in Figure 18, mass spectral analyses of amplification products of 77 base pair bioagent identifying amplicons containing PAG01 SNP loci clearly indicate the capability of using the molecular mass for distinguish the identity of the SNP.

For MLVA analysis, amplification products of bioagent identifying amplicons containing portions of pX01, pX02, vrrB2, CG3, Bavntr12, and Bavntr35 loci were examined. Results are shown in Table 8 and indicate that the present methods are capable of distinguishing the molecular masses and base compositions of the MLVA markers at the six different loci.

Table 8: MLVA VNTR loci information and complete base composition data for observed *B. anthracis* alleles as detected by ES-FTICR-MS

MLVA Marker	Number of Observed Alleles	Size of Allele (base pairs)	Repeat Structure	ESI-FTCR-MS Derived Base Compositions (A:G:C:T)
BaVNTR12	1 2	112 114	AT	39:20:19:35 40:20:19:36
BaVNTR35	1 2 3 4	103 109 115 121		28:13:19:43 29:15:19:46 30:17:19:49 31:19:19:52
pX01	1 2 3 4 5 6 7	119 122 125 128 131 134 143	AAT	51:14:15:40 53:14:15:41 55:14:15:42 57:14:15:43 59:14:15:44 61:14:15:45 67:14:15:48
pX02	1 2 3 4 5 6	135 137 139 141 143 155	AT	37:23:27:48 38:23:27:49 39:23:27:50 40:23:27:51 41:23:27:52 47:23:27:58
CG3	1 2	153 158		61:14:23:55 64:14:23:57
VrrB2	1 2 3	150 159 168		54:21:51:27 59:21:53:29 64:21:55:31

The results presented in this example indicate that the methods of the present invention provide a powerful means for high-throughput genotyping of *B. anthracis*. These same methods can be applied to similar genotyping analyses for other bioagents.

Various modifications of the invention, in addition to those described herein, will be apparent to those skilled in the art from the foregoing description. Such modifications are also intended to fall within the scope of the appended claims. Each reference cited in the present application is incorporated herein by reference in its entirety